STRATEGY RECOVERY FOR STOCHASTIC MEAN PAYOFF GAMES

MARCELLO MAMINO

Abstract. We prove that to find optimal positional strategies for stochastic mean payoff games when the value of every state of the game is known, in general, is as hard as solving such games tout court. This answers a question posed by Daniel Andersson and Peter Bro Miltersen.

In this note, we consider perfect information 0-sum stochastic games, which, for short, we will just call *stochastic games*. For us, a stochastic game is a finite directed graph whose vertices we call *states* and whose edges we call *transitions*, multiple edges and loops are allowed but no state can be a sink. To each state s is associated an *owner* $o(s)$ which is one of the two players Max and Min. Each transition $s \xrightarrow{A,p} t$ has an *action* A and a *probability* $p \in \mathbb{Q} \cap [0,1]$, with the condition that, for each state s, the probabilities of the transitions exiting s associated to the same action must sum to 1. We say that the action A is *available* at state s if one of the transitions exiting s is associated to A. Furthermore to each action A is associated a *reward* $r(A) \in \mathbb{Q}$.

A play of a stochastic game G begins in some state s_0 and produces an unending sequence of states $\{s_i\}_{i\in\mathbb{N}}$ and actions $\{A_i\}_{i\in\mathbb{N}}$. At move i, the owner of the current state s_i chooses an action A_i among those available at s_i, then one of the transitions exiting s_i with action A_i is selected at random according to their respective probabilities, and the next state s_{i+1} is the destination of the chosen transition. A play can be evaluated according to the β-discounted payoff criterion

$$\mathbf{v}_\beta\left(A_0, A_1 \ldots\right) = (1-\beta)\sum_{i=0}^{\infty} r\left(A_i\right)\beta^i$$

for $\beta \in [0,1)$. Or it can be evaluated according to the mean payoff criterion

$$\mathbf{v}_1\left(A_0, A_1 \ldots\right) = \liminf_{n\to\infty} \frac{1}{n+1}\sum_{i=0}^{n} r\left(A_i\right)$$

The goal of Max is to maximize the evaluation, that of Min is to minimize it. It is known that for both criteria there are optimal strategies which are *positional* [Gil57, LL69], namely such that the action chosen at s_i depends only on the state s_i – an not, for instance, on the preceding states in the play, on i, or on a random choice. Given two positional strategies σ and τ for Max and Min respectively, and given $\beta \in [0,1]$, we denote $\mathbf{v}_\beta(G, s_0, \sigma, \tau)$ the expected value of $\mathbf{v}_\beta$ on all

Laboratoire d'Informatique de l'École Polytechnique (lix), Bâtiment Alan Turing, 1 rue Honoré d'Estienne d'Orves, Campus de l'École Polytechnique, 91120 Palaiseau, France.
E-mail address: mamino@lix.polytechnique.fr.
Date: 16·vi·2015.
The author has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 257039).

plays generated by σ and τ starting from s_0. We write $\mathbf{v}_\beta(G, s_0)$ for $\mathbf{v}_\beta(G, s_0, \sigma, \tau)$ with σ and τ optimal. For basic information on stochastic games one may refer to the book [FV97].

Given a stochastic game with probabilities and rewards encoded in binary, and a value of β also encoded in binary, it makes sense to study the computational complexity of the task of solving the game. Strategically solving a game, as defined in [AM09], means to find a pair of optimal strategies. Quantitatively solving G means to find $\mathbf{v}_\beta(G, s)$ for all states s. In general, the second task is easier than the first. The *strategy recovery problem* is, given the quantitative solution of a game, to produce a strategic solution. It has been observed in [AM09] that this task can be performed trivially in linear time for discounted payoff games, and also, but not trivially, for terminal payoff and simple stochastic games, hence it was asked whether the same could be done for stochastic mean payoff games (this is, indeed, the only missing element to complete Andersson and Miltersen's picture). Our aim is to prove that the strategy recovery problem for stochastic mean payoff games is as hard as it possibly can.

Theorem 1. *The strategy recovery problem for stochastic mean payoff games is equivalent, modulo polynomial time Turing reductions, to the task of strategically solving mean payoff games.*

We will combine the reduction from stochastic mean payoff to discounted payoff games proven in [AM09] with a new reduction from discounted to mean payoff games of a special form that we call β-recurrent. Then we will show that β-recurrent mean payoff games can be turned into *strategically equivalent* mean payoff games having the additional property that all states have value 0. For this latter class of games, the strategy recovery problem is obviously equivalent to solving the games strategically.

Definition 2. Let G be a stochastic game and s_0 one of the states of G. We define the β*-recurrent game associated to* G *and* s_0, denoted G_{β,s_0}. The game G_{β,s_0} has the same state-space as G. Each transition $a \xrightarrow{A,p} b$ in G is replaced by two new transitions in G_{β,s_0}, namely $a \xrightarrow{A,\beta p} b$ and $a \xrightarrow{A,(1-\beta)p} s_0$. The first of these new transitions will be called *of the first kind*, the second *of the second kind*. We say that a game is β*-recurrent* if it results from the construction just defined, for some G.

Notice that our β-recurrent games are ergodic in the sense of [BEGM10]. The complexity of ergodic games has been settled in a recent work [CIJ14a] (see the full version [CIJ14b]), however we need for our reduction the extra properties of β-recurrent games. Interestingly, the definition of ergodic in [CIJ14a] is more restrictive than that in [BEGM10], and, in particular, in this stronger sense, a β-recurrent game may not be ergodic, nor an ergodic game needs to be β-recurrent.

Lemma 3. *The task of quantitatively solving stochastic discounted payoff games is polynomial time Turing reducible to quantitatively solving* β*-recurrent stochastic mean payoff games.*

Proof. Consider a stochastic game G and discount factor β. Let s_0 denote a state of G. We will show that

$$\mathbf{v}_\beta\left(G, s_0\right) = \mathbf{v}_1\left(G_{\beta,s_0}, s_0\right)$$

Intuitively, an infinite play of G_{β,s_0} can be seen as a sequence of finite sub-plays, each of which lasts until a transition of the second kind is taken and the game is reset to the initial state s_0. Each sub-play lasts at least one move, but a second move is played only with probability β, a third one with probability β^2, and so on, thus imitating the discounted payoff situation.

In order to prove the proposition, it suffices to show that, for any pair of positional strategies σ and τ for MAX and MIN respectively, one has

$$(\star) \qquad \mathbf{v}_1\left(G_{\beta,s_0}, s_0, \sigma, \tau\right) = \mathbf{v}_\beta\left(G, s_0, \sigma, \tau\right)$$

In fact, it follows from this equation that σ and τ are a pair of optimal positional strategies for G_{β,s_0} if and only if they are a pair of optimal positional strategies for G with starting position s_0.

It remains to prove equation $(\star)$. For each state s of G, call $A_{\sigma,\tau}(s)$ the action chosen by either σ or τ (according to the owner of s) at the state s. The β-discounted values of the states of G are determined by the condition

$$\mathbf{v}_\beta\left(G, s, \sigma, \tau\right) = (1-\beta)\, r\left(A_{\sigma,\tau}\left(s\right)\right) + \sum_{t\in G} \beta p_{\sigma,\tau}\left(s \to t\right) \mathbf{v}_\beta\left(G, t, \sigma, \tau\right)$$

where $p_{\sigma,\tau}(v \to w)$ denotes the probability that, from state s, a transition to state t is chosen when playing strategy σ against τ. If we call $s_0 \dots s_n$ the states of G and $\bar{v}_\beta = (\mathbf{v}_\beta(G, s_i, \sigma, \tau))_{i=1\dots n}$ the value vector of G, then the condition above can be rewritten in the form

$$\bar{v}_\beta = (1-\beta)\, \bar{r} + \beta P \bar{v}_\beta$$

where $\bar{r}$ is the vector of the rewards $\bar{r}_i = r(A_{\sigma,\tau}(s_i))$, and P denotes the matrix of the transition probabilities $P_{i,j} = p_{\sigma,\tau}(s_i \to s_j)$. Hence

$$\bar{v}_\beta = (1-\beta)\, (I - \beta P)^{-1}\, \bar{r}$$

where I denotes the $n \times n$ identity matrix.

Now we turn our attention to the mean payoff of the pair of strategies σ and τ in G_{β,s_0}. We can compute $\mathbf{v}_1(G_{\beta,s_0}, s_0, \sigma, \tau)$ averaging the rewards over the stable distribution of the Markov chain induced by these strategies on the states of G. This stable distribution μ must be unique, because, by virtue of G_{β,s_0} being β-recurrent, the Markov chain is connected. Moreover μ is determined by the condition

$$\mu\left(s\right) = (1-\beta)\, \delta_{s_0}\left(s\right) + \sum_{t\in G} \beta p_{\sigma,\tau}\left(t, s\right) \mu\left(t\right)$$

where $\delta_{s_0}(s)$ is 1 if $s = s_0$ and 0 otherwise. Rewriting as above, we get

$$\bar{\mu} = (1-\beta)\, e_0 + \beta P^T \bar{\mu}$$

where e_0 is the first element of the canonical basis and $\bar{\mu}_i = \mu(s_i)$. Hence

$$\bar{\mu} = (1-\beta) \left(I - \beta P^T\right)^{-1} e_0$$

Now, computing the average

$$\begin{aligned}
\mathbf{v}_1\left(G_{\beta,s_0}, s_0, \sigma, \tau\right) &= \sum_{s\in G} \mu(s)\, r\left(A_{\sigma,\tau}(s)\right) \\
&= \bar{\mu}^T \bar{r} \\
&= e_0^T (1-\beta)(I-\beta P)^{-1}\, \bar{r} \\
&= e_0^T \bar{v}_\beta \\
&= \mathbf{v}_\beta\left(G, s_0, \sigma, \tau\right)
\end{aligned}$$

□

Lemma 4. *The task of strategically solving β-recurrent stochastic mean payoff games is polynomial time many-one reducible to the strategy recovery problem for stochastic mean payoff games.*

Proof. Let G_{β,s_0} be a β-recurrent stochastic game. As we noticed, all the states of G_{β,s_0} have the same value. Nevertheless, we have no obvious way to determine this value in order to complete the reduction. Instead, we choose to construct a new mean payoff game G' in such a way that all the states of G' get mean payoff value equal to 0, and nonetheless a pair of optimal strategies for G_{β,s_0} can be recovered from a pair of optimal strategies for G'. This is clearly sufficient to establish the lemma.

The game G' is constructed as two chained copies G^1 and G^2 of G_{β,s_0}, redirecting all the transitions of the second kind in each instance – that go to the state corresponding to s_0 in that instance – to the s_0-state in the other. The states of G^1 have the same owner as in G_{β,s_0}, and the transitions originating in G^1 are associated to the same actions with the same rewards as in G_{β,s_0}. In G_2, however, the owners are switched and the signs of the rewards exchanged (formally we replace each action A with a new one A' having $r(A') = -r(A)$). If both players play optimally, we may expect each to win in G_1 precisely as much as he loses in G_2, hence, arguably the value of G' should be 0. On the other hand, in order to play optimally in G', one should play optimally in both the components, so we should be able to extract optimal positional strategies for G_{β,s_0} from optimal positional strategies for G' by mere restriction to the component G^1. We will now proceed to prove our statement.

Let us denote by s^1 and s^2 respectively the states of G^1 and G^2 corresponding to a given state s of G_{β,s_0}. First observe that a play of G', almost surely, will eventually reach state s_0^1, from this follows that all the states of G' must have the same value (G' is ergodic). A positional strategy σ for Max in G' can be seen as a pair of positional strategies (σ^1, σ^2) where σ^1 is the strategy for Max in G_{β,s_0} that we get restricting σ to G^1, and σ^2 is the strategy for Min in G_{β,s_0} that we get from the restriction of σ to G^2 (remember that in G^2 the players are switched). Similarly a strategy τ for Min in G' can be seen as a pair of strategies (τ^1, τ^2) in G_{β,s_0}, the first one for Min and the second for Max. We will prove that for any σ and τ

$$(\star\star)\qquad \mathbf{v}_1\left(G', \cdot, \sigma, \tau\right) = \frac{1}{2}\mathbf{v}_1\left(G, \cdot, \sigma^1, \tau^1\right) - \frac{1}{2}\mathbf{v}_1\left(G, \cdot, \tau^2, \sigma^2\right)$$

From this equation, it follows at once that σ is an optimal strategy for G' if and only if (σ^1, σ^2) is a pair of optimal strategies for G_{β,s_0}, and, in particular, the value of G' is 0.

We turn now to the proof of equation ($\star\star$). Consider the unique stable distribution μ of the Markov process induced by σ and τ. Observe that, independently from σ and τ, at any given state, our Markov chain has probability β of transitioning to a state belonging to the same component, and probability $1-\beta$ of switching component. It follows that the sequence of the components must obey the law of a two-state Markov chain with transition matrix

$$\begin{pmatrix} \beta & 1-\beta \\ 1-\beta & \beta \end{pmatrix}$$

Hence $\mu(G^1) = \mu(G^2) = 1/2$. It suffices to prove that the probability distributions μ^1 and μ^2 defined on the states of G_{β,s_0} by $\mu^1(s) = 2\mu(s^1)$ and $\mu^2 = 2\mu(s^2)$ are the stable distributions induced on G_{β,s_0} by the pairs of strategies (σ^1, τ^1) and (τ^2, σ^2) respectively.

By symmetry, we can concentrate on μ^1. Let $p_{\sigma,\tau}(t,s)$ denote the probability of the transition $t \to s$ in the Markov process induced by the strategies σ and τ. Since all states of G^1 except s_0^1 are only reachable from within G^1 itself, the consistency equation for μ being a stable distribution on G'

$$\mu(s) = \sum_{t \in G'} p_{\sigma,\tau}(t,s)\, \mu(t)$$

implies the same condition for μ^1 at all states except s_0. At s_0 one concludes by direct computation observing that the component of the sum on the right hand side due to transitions of the second kind must be

$$(1-\beta)\,\mu\left(G^2\right) = \frac{1-\beta}{2} = (1-\beta)\,\mu\left(G^1\right) \qquad \square$$

Proof of Theorem 1. By [AM09, Theorem 1], solving stochastic mean payoff games strategically is reducible to solving stochastic discounted payoff games quantitatively, which reduces, by Lemma 3, to solving β-recurrent stochastic mean payoff games quantitatively. In turn, solving such β-recurrent games quantitatively is reducible to solving the same strategically, just because they are, in particular, stochastic mean payoff games. By Lemma 4, this final task is reducible to the strategy recovery problem for stochastic mean payoff games. $\square$

Finally, we would like to remark that our construction relies on the interpretation of *strategic solution* as requiring optimal *positional* strategies. Were a more general class of strategies available, then the problem of finding an optimal one would become easier. In particular, the games produced by Lemma 4 happen to be symmetric under switching the players and the signs of the rewards. Under this circumstance, it would not be surprising if one could play optimally by some form of strategy stealing technique.

Acknowledgements

We would like to express gratitude to Manuel Bodirsky and Eleonora Bardelli for interesting discussions.

References

[AM09] Daniel Andersson and Peter Bro Miltersen. The complexity of solving stochastic games on graphs. In *Algorithms and computation*, volume 5878 of *Lecture Notes in Comput. Sci.*, pages 112–121. Springer, Berlin, 2009.

[BEGM10] Endre Boros, Khaled Elbassioni, Vladimir Gurvich, and Kazuhisa Makino. A pumping algorithm for ergodic stochastic mean payoff games with perfect information. In *Integer programming and combinatorial optimization*, volume 6080 of *Lecture Notes in Comput. Sci.*, pages 341–354. Springer, Berlin, 2010.

[CIJ14a] Krishnendu Chatterjee and Rasmus Ibsen-Jensen. The complexity of ergodic mean-payoff games. In Javier Esparza, Pierre Fraigniaud, Thore Husfeldt, and Elias Koutsoupias, editors, *Automata, Languages, and Programming*, volume 8573 of *Lecture Notes in Computer Science*, pages 122–133. Springer Berlin Heidelberg, 2014.

[CIJ14b] Krishnendu Chatterjee and Rasmus Ibsen-Jensen. The complexity of ergodic mean-payoff games. *CoRR*, abs/1404.5734, 2014.

[FV97] Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer-Verlag, New York, 1997.

[Gil57] Dean Gillette. Stochastic games with zero stop probabilities. In *Contributions to the theory of games, vol. 3*, Annals of Mathematics Studies, no. 39, pages 179–187. Princeton University Press, Princeton, N. J., 1957.

[LL69] Thomas M. Liggett and Steven A. Lippman. Stochastic games with perfect information and time average payoff. *SIAM Rev.*, 11:604–607, 1969.